Oponentský posudek diplomové práce

Michaela Pilátová: Filtrování informací v XML dokumentech

Mezi cíle této práce patřilo:

- popsat stávající algoritmy pro filtrování XML dokumentů,
- na základě studia navrhnout vlastní filtrovací systém,
- implementovat prototypovou verzi návrhu,
- ověřit chování navrženého systému na vhodné kolekci dat.

Textová část práce je rozdělena do několika kapitol, které odpovídají těmto cílům. První kapitoly obsahují především popis základů jazyka XML, jazyků DTD a XML Schema pro popis jejich schémat, a dotazovacího jazyka XPath. Zmíněny jsou i způsoby reprezentace a parsování XML dokumentů.

Jádro popisné části práce tvoří pátá kapitola. Ta obsahuje popisy algoritmů XFilter a Yfilter pro filtrování dokumentů, a dále popis obecné struktury filtrovacích systémů.

Za hlavní přínos práce lze označit kapitolu šestou, která obsahuje popis vlastního algoritmu distribuce uživatelských profilů, který vychází z popsaných algoritmů a zlepšuje rozložení zátěže mezi jednotlivé uzly systému vhodnějším výběrem bodů rozpadu prefixového stromu dotazů.

Po formální stránce je text pěkně členěný a souvisle čitelný. Narazil jsem jen na zanedbatelné množství chyb a překlepů. Zavádějící je např. příklad Xpath výrazu v kapitole 4.3.4 na str. 31, kde je u výrazu „`//cena[@dph='5'][2]`“ uvedeno, že vrací druhého zaměstnance z dokumentu s id rovným hodnotě 101.

Prototypová implementace, která je součástí práce, je úzce zaměřena především na problematiku filtrování informace. Dovoluje zvolit jednu z více v práci popisovaných strategií filtrování, a tak porovnávat výsledky navrženého modelu se stávajícími. Prováděné texty a jejich výsledky jsou uváděny v kapitole 6.6. Je zde uvedeno, že testy byly prováděny na 2000 vygenerovaných dotazech. Na CD jsem však dohledal pouze jednoduchý seznam s pěti dotazy, a tak bych se rád zeptal, nakolik byly dotazy nad použitým DTD různé. Kromě výsledků uvedených v práci by bylo zajímavé vědět, jak závisí časová náročnost zpracování dokumentu právě na počtu různých dotazů evidovaných v systému. Už proto, že v práci se na několika místech hovoří o tom, že podobné systémy mohou mít i miliony uživatelů. Jak je v práci uvedeno, implementace organizuje jednotlivé brokery do hvězdy s jedním centrálním kořenovým brokerem, který rozděluje dokumenty ostatním tak, aby neposílal dokumenty či jejich části, které nejsou na cílovém brokeru potřebné. Zde se nabízí otázka, nakolik je zjednodušení topologie brokerů dáno navrženým algoritmem. Tj. je možné algoritmus použít i v případě, že budou brokery uspořádány do obecnějšího stromu? Navržené prototypové řešení obsahuje dva centrální uzly, jejichž kolaps vyřadí z provozu celý systém – koordinátora a kořenový broker. Nebylo by jednodušší, aby roli koordinátora hrál přímo kořenový broker, respektive, aby ve víceúrovňové architektuře každý broker koordinoval ty pod ním?

Celkově se domnívám, že přes výše uvedené připomínky předkládané řešení naplnilo zadání a splňuje požadavky kladené na diplomové práce. Navrhuji proto práci uznat jako práci diplomovou.

V Praze dne 26. 1. 2006

RNDr. Michal Kopecký, Ph.D.
KSI MFF UK